Collecting, Classifying, Analyzing, and Using Real-World Elections

Niclas Boehmer and Nathan Schaar

Algorithmics and Computational Complexity, TU Berlin, Germany

Properties of our data
- virtually no ties
- various sizes
- even complete elections are "large"

$$\text{raw} \xrightarrow{\text{delete vot.+cand.}} \text{complete} \xrightarrow{\geq 15 \text{ cand.}} \text{relevant complete}$$

Status Quo in Computational Social Choice

- Computational Social Choice: algorithmic and axiomatic analysis of collective decision-making problems, where the preferences of agents should be aggregated into a "compromise" solution.
- Early years: study of the theoretical worst-case computational complexity of decision-making related problems.
- More recently: focus has partially shifted towards the practical applicability of theoretical research, yet many subareas still lack empirical research (one explanation: unavailability of real-world data).
- PrefLib platform largest database for real-world elections (previously containing 701 real-world elections divided into 36 datasets). Most of them either have few candidates or voters express only partial preferences which can include many ties.
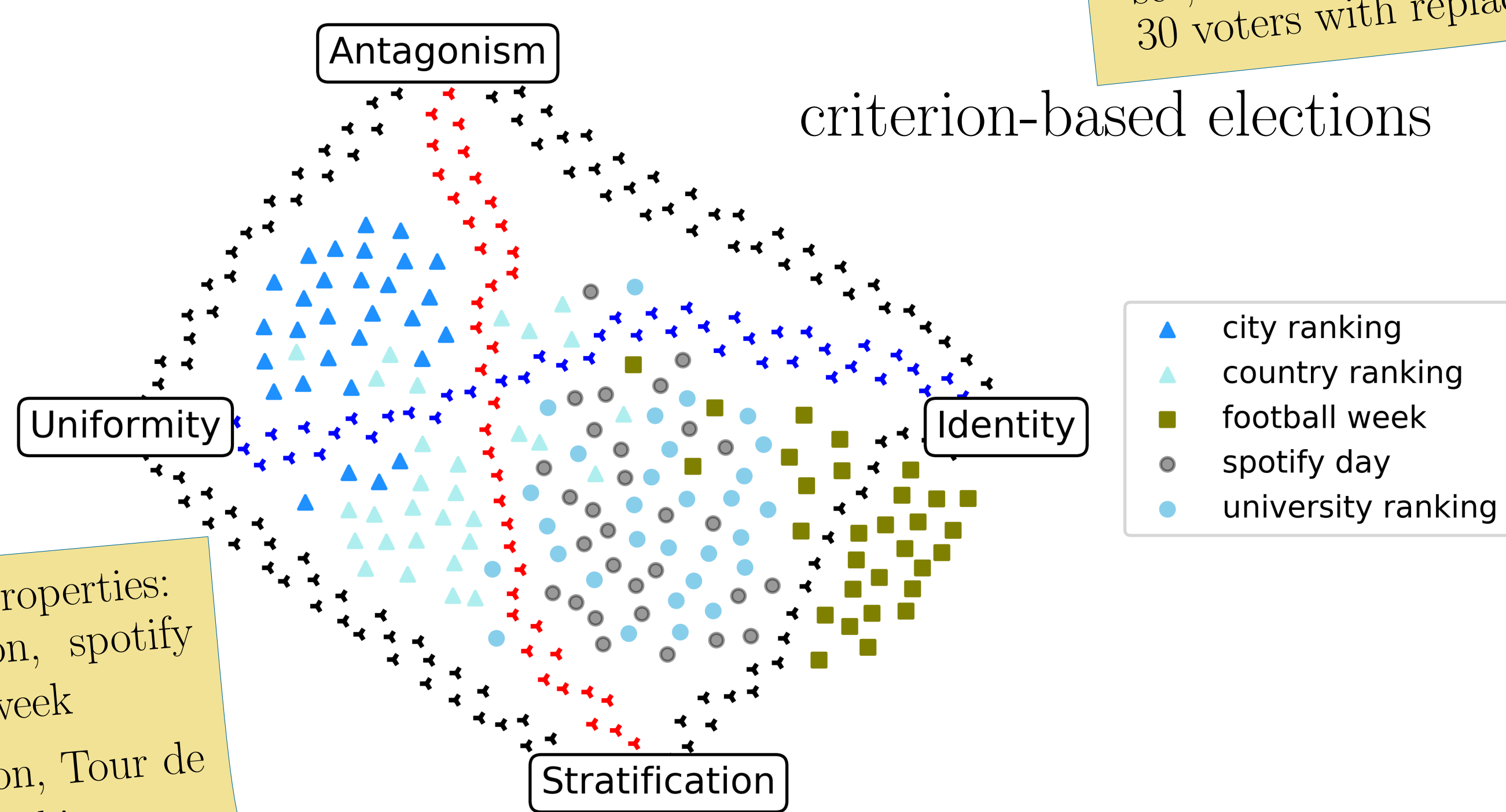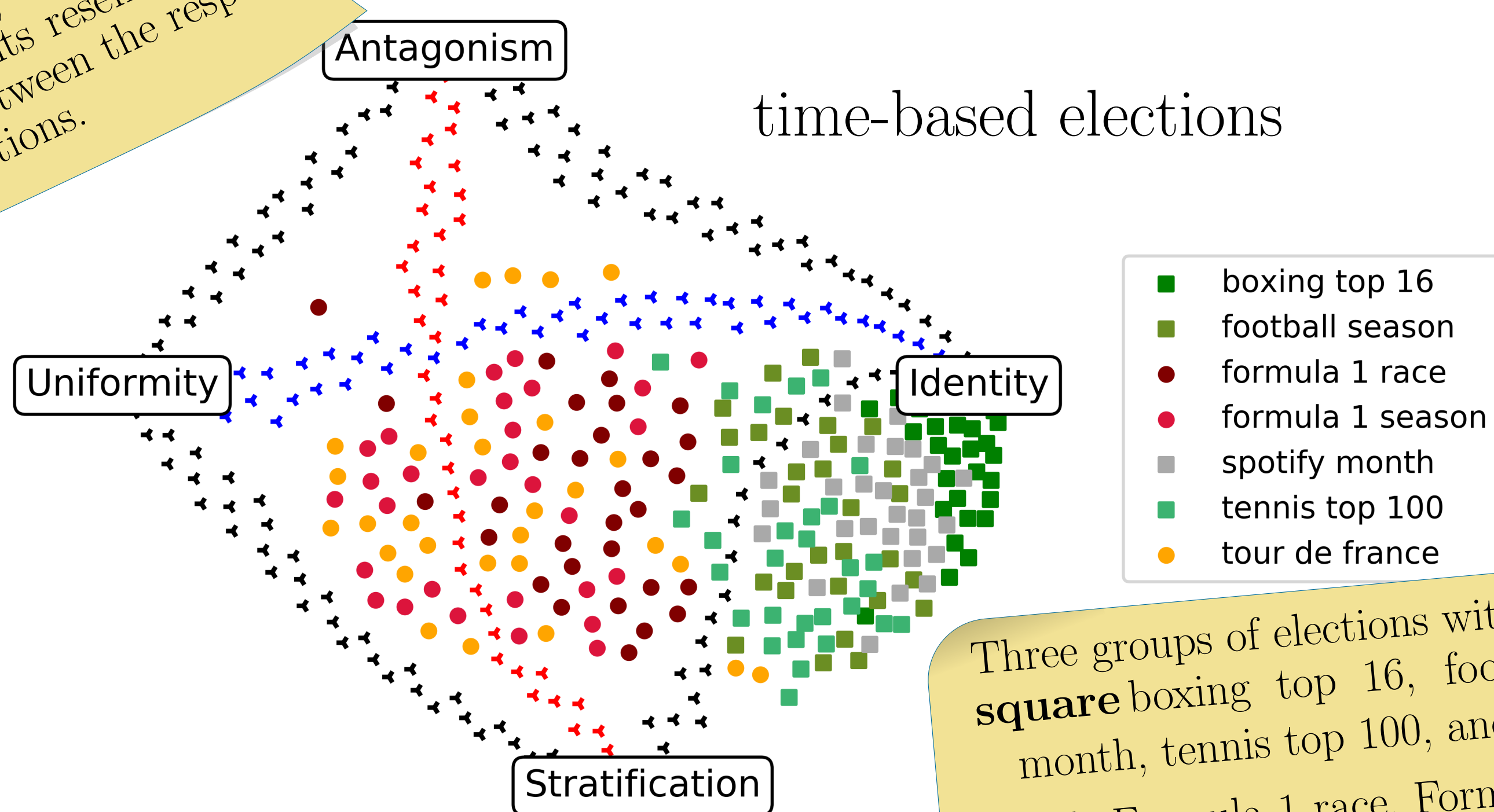
Collecting Elections (extract)

name	type	raw			relevant complete		
		#Elec.	Avg. #Voters	Avg. #Cand.	#Elec.	Avg. #Voters	Avg. #Cand.
boxing top 16	time	99	31.9	19.76	31	17.45	15.32
football season	time	2746	12.28	152.36	2422	12.6	156.71
Formula 1 race	time	454	61.3	20.46	396	47.2	17.93
Formula 1 season	time	71	14.58	43.97	42	13.38	21.57
spotify month	time	645	29.78	306.64	632	29.91	109.28
tennis top 100	time	29	50.48	140	29	49.9	62.31
Tour de France	time	97	21.14	175.69	95	19.7	82.64
city ranking	crit.	1	12	216	1	12	216
country ranking	crit.	12	17.25	119.17	12	14.25	95.58
football week	crit.	415	83.28	219.67	415	77.35	98.45
spotify day	crit.	362	53.06	2427.74	357	49.06	20.73
university ranking	crit.	4	18.5	832.5	4	18.5	123.25

⇒ 7582 real-world elections divided into 25 datasets available at tinyurl.com/real-elections.

For our experiments, we used *normalized* elections: For each dataset we drew 500 elections, then took a 15-candidates subset, and subsequently sampled 30 voters with replacement.

Classifying Elections — A Map of Real-World Elections

Each point corresponds to an election; distances between points resemble distances between the respective elections.

time-based elections

Legend: boxing top 16, football season, formula 1 race, formula 1 season, spotify month, tennis top 100, tour de france

criterion-based elections

Legend: city ranking, country ranking, football week, spotify day, university ranking

Three groups of elections with similar properties:
square boxing top 16, football season, spotify month, tennis top 100, and football week
circle Formula 1 race, Formula 1 season, Tour de France, spotify day, and university rankings
**triangle** city and country rankings

Analyzing Elections — Similarity Scores

	max. dist.	average dist.	disagr. pairs	Kemeny score
boxing top 16	10.39	4.40	11.43	97.20
football season	24.94	11.78	34.84	251.37
formula 1 race	71.61	32.76	99.45	698.09
formula 1 season	62.69	36.59	96.45	814.77
spotify month country	20.94	8.86	27.72	187.46
tennis top 100	32.68	15.41	43.87	336.30
tour de france	65.91	37.78	96.44	855.50
city ranking	92.69	49.17	104.93	1203.46
country ranking	85.00	43.82	104.73	1004.67
football week	32.98	14.42	50.69	298.41
spotify day world	61.08	32.05	89.80	703.80
university ranking	66.25	32.27	95.25	691.09
aggregated	52.26	26.61	71.30	595.18

Take-aways

- For most elections similarity measures are not small.
- Datasets quite homogeneous with respect to similarity of votes.
- Kemeny score highly correlated with the average swap distance.

Analyzing Elections — Restricted Domains

(a) All elections   (b) spotify month elections   (c) spotify day elections
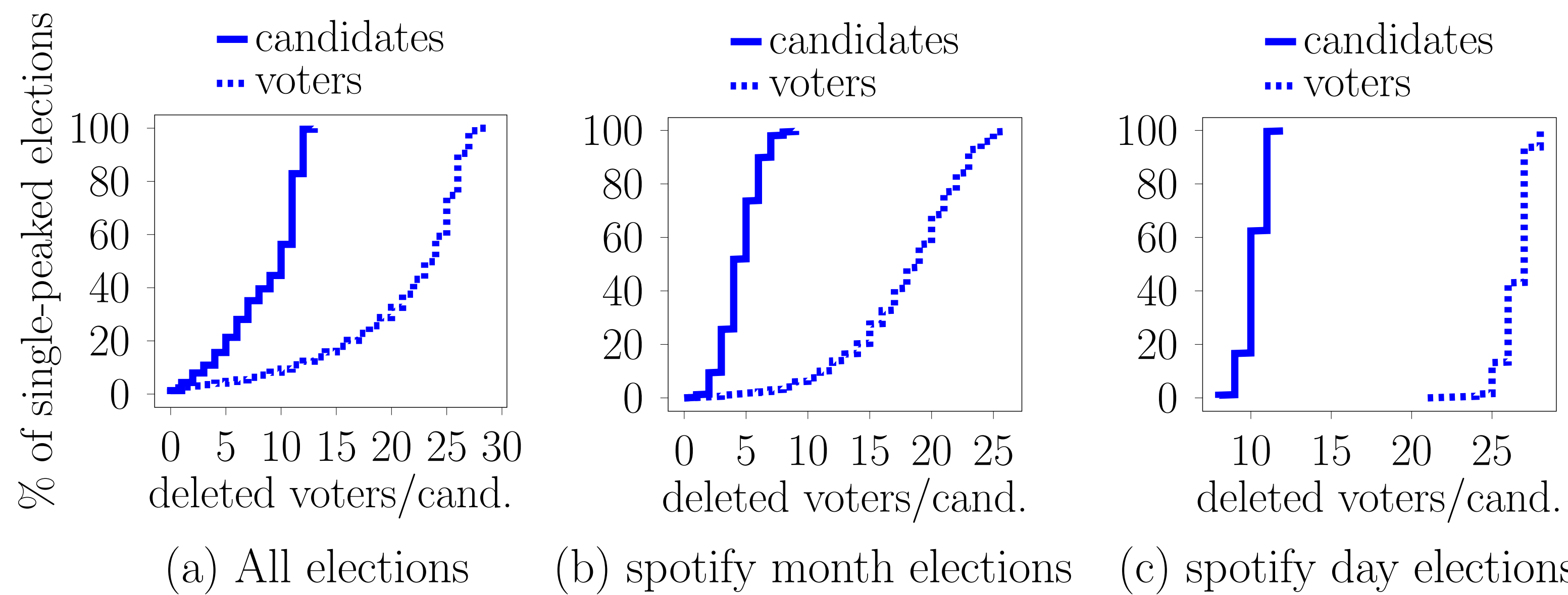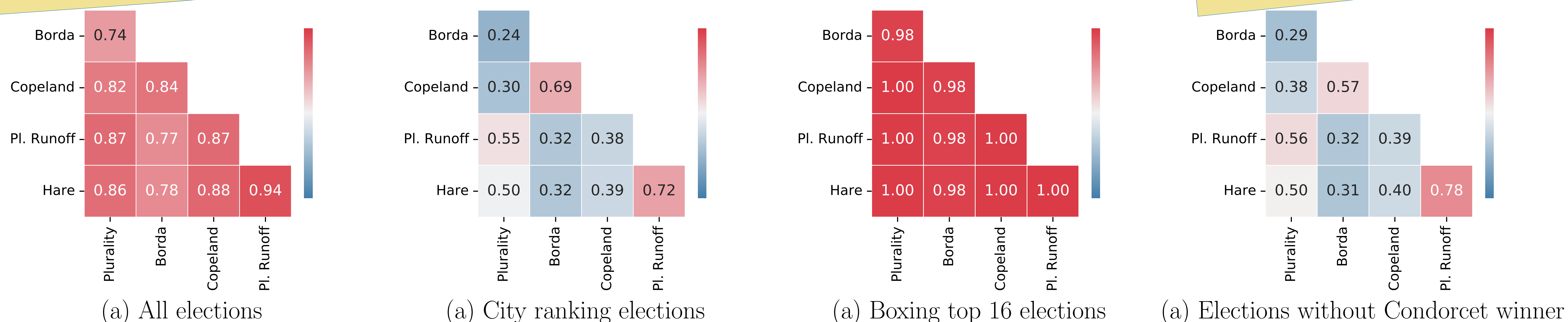
Take-aways

- There are only few elections from a restricted domain and only some elections close to one.
- Elections that are close to one domain are typically also close to another.
- Elections from a restricted domain are typically quite degenerate.

Most of our elections (86%) have a Condorcet winner and all voting rules often select them as a winner (>88%).

single-peaked ⤳ there is a societal candidate order and each voter ranks candidates that are closer to its top-choice according to the order above further away ones.

Using Elections — Similarity of Voting Rules

(a) All elections   (a) City ranking elections   (a) Boxing top 16 elections   (a) Elections without Condorcet winner

Takeaway: Voting rules often agree on the returned winner because most elections have a Condorcet winner and voting rules often select them.